PROCEDE D'ARCHIVAGE DE TEXTES ET PROCEDE DE RECHERCHE, PARMI LES TEXTES ARCHIVES, DE CEUX QUI SONT PERTINENTS A L'EGARD D'UNE QUESTION

Avec les moyens modernes de traitement de données, d'une rare permanence, le monde de la documentation a connu récemment une expansion considérable. Au fur et à mesure que les besoins ou les envies de connaissances des individus augmentent, les données elles-mêmes augmentent également, peut-être davantage encore. Le nombre d'ouvrages, revues, journaux, et autres publications de toutes sortes, même sur une question précise, ne fait que croître aussi. Le stockage ou l'archivage de données est devenu une tâche difficile. A l'inverse, l'extraction de données d'un lot stocké n'est, aujourd'hui, pas plus aisé.

On connaît la solution des mots clés à ce double problème. Compte tenu des tailles des banques de données, c'est une solution qui, bien souvent, n'est plus appropriée, l'interrogation d'un mot clé fournissant à la fois trop et pas assez de documents, en raison des homonymies (documents non pertinents) et des synonymies non prises en compte.

De microscopiques, l'analyse et la recherche doivent devenir macroscopiques et c'est ce que la demanderesse a cherché à offrir. Du mot, les documentalistes et archivistes doivent passer au concept, à l'idée, c'est-à-dire à la pluralité, la combinaison, l'association de mots.

L'invention concerne aussi bien le processus d'analyse et d'archivage-stockage de textes, que la recherche, l'extraction de textes archivés. Bref, l'invention vise à proposer des outils d'amélioration de l'organisation des connaissances.

L'invention concerne tout d'abord un procédé d'archivage d'un texte selon lequel:
- on crée dans un repère conceptuel multidimensionnel un dictionnaire de mots,
- on compare chaque mot conceptuel d'au moins une portion du texte à archiver à ceux du dictionnaire pour déterminer la position de ce mot dans ledit repère et
- on détermine la résultante des positions de tous les mots conceptuels de la portion de texte à archiver pour déterminer la position d'une

conceptualisation globale de la portion de texte dans ledit repère et archiver cette position.

Par le terme "mot", il faut naturellement entendre l'unité linguistique, c'est-à-dire le mot, au sens propre du terme, mais également le groupe de mots formant une expression sémantique unitaire, comme par exemple "crise cardiaque".

Les axes du repère de l'invention, en nombre égal à celui des dimensions, correspondent aux divers concepts exprimés dans le dictionnaire.

Un mot, dans le procédé de l'invention, est défini par un point, ou un vecteur s'étendant depuis l'origine du repère jusqu'à ce point, dont les coordonnées, sur les axes du repère, correspondent respectivement aux poids relatifs des divers concepts attachés à ce mot.

Finalement, le procédé d'archivage de l'invention consiste à vectoriser les mots d'un texte et à en calculer la résultante conceptuelle représentative de l'ensemble du texte dans un repère d'une pluralité de concepts.

Avantageusement, pour déterminer la résultante des positions, dans le repère, de tous les mots conceptuels de la portion de texte à archiver, à chaque position de mot dans le repère, on associe d'abord sa position dans le texte et son rôle syntaxique.

Avantageusement encore, pour déterminer la résultante des positions des mots conceptuels de la portion de texte à archiver, on multiplexe ces positions par un algorithme de composition.

L'invention concerne aussi un procédé de recherche parmi une pluralité de textes archivés selon le procédé ci-dessus, de ceux qui traitent d'une question, dans lequel:
- comme pour l'archivage de texte, on détermine la position dans le repère conceptuel multidimensionnel d'une conceptualisation globale de la question, par détermination de la résultante des positions de tous les mots conceptuels de la question et
- on compare la position de la conceptualisation globale de la question aux positions homologues des textes archivés, pour retenir au moins l'une d'entre elles, correspondant à un texte recherché.

Avantageusement, on compare les positions des conceptualisations globales de la question et des textes archivés en déterminant, pour chaque texte, une distance entre les deux positions respectives de la question et du texte.

De préférence, la distance déterminée entre deux positions est non euclidienne.

L'invention sera mieux comprise à l'aide de la description suivante de différents modes de réalisation du procédé d'archivage de textes et du procédé de recherche, parmi les textes archivés, de ceux qui traitent d'une question, en référence à la figure unique annexée qui représente un repère conceptuel multidimensionnel.

Par souci de clarté, et de façon à faire comprendre au mieux l'invention, l'exemple qui va maintenant être décrit est un exemple didactique, un cas d'école, extrêmement simplifié.

Le procédé d'archivage de textes va d'abord être explicité.

1- Procédé d'archivage de textes

1.1- Création d'un dictionnaire de mots

D'emblée, on rappelle que par le terme "mot", on entend désigner une unité linguistique, c'est-à-dire aussi bien un mot, au sens propre du terme, qu'un groupe de mots formant une expression sémantique unitaire, comme par exemple "crise cardiaque", "carte d'identité", "secteur secondaire", etc..

Soit un espace vectoriel de dimension n, n étant un entier naturel supérieur à un, que l'on munit d'un repère conceptuel $\mathfrak{R}$, d'un produit scalaire et d'une norme associée. On rend le repère $\mathfrak{R}$ orthonormé. Par repère orthonormé, on entend désigner une base de n vecteurs orthogonaux (pour le produit scalaire défini) et de norme égale à un (pour la norme définie). Par définition, les vecteurs de la base sont des vecteurs par combinaison linéaire desquels tous les vecteurs de l'espace vectoriel peuvent être définis.

Dans l'exemple didactique de la description, l'espace vectoriel est de dimension trois et muni du produit scalaire euclidien et de la norme euclidienne associée, ainsi que d'un repère conceptuel $\mathfrak{R}$, représenté sur la

figure, comportant trois axes A_1, A_2, A_3 portant respectivement les vecteurs de base $\vec{u}_1$, $\vec{u}_2$, $\vec{u}_3$ dont les coordonnées respectives dans le repère $\mathfrak{R}$ sont (1, 0, 0), (0, 1, 0) et (0, 0, 1).

D'emblée, on notera qu'une position dans le repère $\mathfrak{R}$ est définie par un triplet de coordonnées respectivement suivant les axes A_1, A_2 et A_3, et qu'à chaque position dans le repère $\mathfrak{R}$ correspond un vecteur de mêmes coordonnées, s'étendant depuis une origine O du repère $\mathfrak{R}$. Par la suite, on confondra donc les termes "position" et "vecteur".

Par définition, le produit scalaire euclidien de deux vecteurs $\vec{X}$ et $\vec{Y}$ est égal à la somme des produits des coordonnées homologues des vecteurs $\vec{X}$ et $\vec{Y}$. La formule mathématique pour le calcul du produit scalaire euclidien est donc la suivante:

$$\langle \vec{X}, \vec{Y} \rangle = \sum_{i=1}^{n} x_i \cdot y_i$$

- $<\vec{X},\vec{Y}>$ représente le produit scalaire de X et de Y et
- x_i et y_i représentent les coordonnées respectives du vecteur X et du vecteur Y suivant l'axe A_i,

avec n représentant la dimension de l'espace vectoriel, égal à trois dans l'exemple de la description.

La norme euclidienne $\|\vec{X}\|$ du vecteur $\vec{X}$ est définie par la formule suivante:

$$\|\vec{X}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

L'unité de chaque axe correspond à un concept, une idée exprimée dans le dictionnaire. En l'espèce:
- l'unité de l'axe A_1 correspond au concept de la physique,
- l'unité de l'axe A_2 correspond au concept de l'état liquide et
- l'unité de l'axe A_3 correspond au concept de l'imprimerie.

La physique, l'état liquide et l'imprimerie sont donc les trois concepts du repère conceptuel $\mathfrak{R}$, correspondant aux trois dimensions du repère $\mathfrak{R}$.

Afin de créer le dictionnaire de mots, on prend, parmi les mots du langage, les mots conceptuels et, pour chacun de ces mots, on détermine sa position dans le repère conceptuel $\mathfrak{R}$.

Par les termes "mot conceptuel", on entend désigner un mot important du texte, chargé de sens, exprimant une ou plusieurs idées et contribuant par conséquent, de façon majeure, à donner au texte son sens global. Bref, un mot conceptuel est un mot susceptible de faire référence à au moins un concept du repère conceptuel.

Par souci de clarté, on crée ici un dictionnaire contenant les seuls mots nécessaires à la compréhension de l'exemple particulier de la description, à savoir les mots suivants: corps, plonger, liquide, subir, poussée, vertical, police, penser, noyade, style, fluide, idée, manquer, mécanique.

On sait qu'un mot peut avoir plusieurs sens et on peut généralement déterminer le sens dans lequel ce mot est employé dans un texte, suivant le contexte du texte.

Pour introduire chacun de ces mots dans le dictionnaire, on recherche tous les sens possibles du mot, on en déduit tous les concepts relatifs au repère $\mathfrak{R}$ auxquels ce mot est susceptible de faire référence, selon le contexte dans lequel il est employé, et, suivant ces concepts, on attribue au mot une position dans le repère conceptuel $\mathfrak{R}$. Les coordonnées de la position de chaque mot correspondent aux poids relatifs des divers concepts attachés à ce mot. Dans le dictionnaire, les mots sont chacun associés à une position représentée par un triplet de coordonnées dans le repère $\mathfrak{R}$.

Afin d'illustrer cette étape de création du dictionnaire, explicitons plus en détails l'introduction de quelques mots particuliers dans le dictionnaire.

Prenons d'abord le mot "corps". D'après le dictionnaire "Le Petit Robert" (édition les dictionnaires Le Robert, 1993), un corps peut désigner "tout corps matériel caractérisé par ses propriétés physiques", et le "corps d'une lettre" s'entend de "la dimension d'un caractère d'imprimerie". On en déduit que le mot "corps" peut, suivant son emploi, faire référence soit au concept de la physique soit au concept de l'imprimerie. En revanche, dans aucun de ses sens, le corps ne fait référence au concept de l'état liquide. Le mot corps est ainsi susceptible de faire référence au concept de la physique (axe A_1)

ainsi qu'à celui de l'imprimerie (axe A_3). En conséquence, on lui attribue, dans le repère conceptuel $\mathfrak{R}$, une position ayant pour coordonnées (1, 0, 1).

Prenons encore le mot "plonger" qui peut notamment signifier "faire entrer dans un liquide", d'après le dictionnaire Le Petit Robert. Ce mot est donc susceptible de faire référence au concept de l'état liquide (axe A_2) mais ne fait référence, dans aucun de ses sens, au concept de la physique (axe A_1) ou à celui de l'imprimerie (axe A_3). Par conséquent, on attribue au mot "plonger" une position ayant pour coordonnées (0, 1, 0) dans le repère conceptuel $\mathfrak{R}$.

Le tableau 1 contient les coordonnées des positions de tous les mots du dictionnaire, déterminées suivant les étapes que l'on vient de détailler pour deux exemples particuliers.

Tableau 1

	Coordonnées		
Mots	A_1	A_2	A_3
corps	1	0	1
plonger	0	1	0
liquide	1	1	0
subir	0	0	0
poussée	1	0	0
vertical	0	0	0
police	0	0	1
penser	0	0	0
noyade	0	1	0
style	0	0	1
fluide	1	1	0
idée	0	0	0
manquer	0	0	0
mécanique	1	0	0

1.2- Conceptualisation globale des textes à archiver

Dans l'exemple didactique de la description, on dispose de trois textes à archiver qui sont les suivants:

Texte 1: "Tout corps plongé dans un liquide subit une poussée verticale."
Texte 2: "La police pense à une noyade."
Texte 3: "Le style est fluide mais les idées manquent."

Dans une étape préalable, on procède à une analyse syntaxique de chaque texte à archiver afin d'en extraire les mots conceptuels.

Grâce à l'extraction des mots conceptuels, on s'affranchit, en vue de l'étape suivante de "vectorisation" du texte, des mots contribuant de façon mineure à donner au texte son sens global, tels que notamment les pronoms, les articles, les prépositions, etc..

Pour illustrer cette étape d'extraction, appliquons-la au texte 1. Après analyse de ce texte et extraction des mots conceptuels, on obtient les mots conceptuels suivants: corps, plongé, liquide, subit, poussée et verticale.

On transforme ensuite les mots conceptuels fléchis (c'est-à-dire les verbes conjugués, les adjectifs accordés, les noms au pluriel, etc.), dans leur forme non fléchie.

Les mots conceptuels extraits des textes 1, 2 et 3, et éventuellement transformés dans leur forme non fléchie, sont répertoriés dans le tableau 2.

Tableau 2

Textes	Mots extraits
1	corps, plonger, liquide, subir, poussée, vertical
2	police, penser, noyade
3	style, fluide, idée, manquer, mécanique

Pour chaque texte à archiver, on détermine la position de chacun des mots conceptuels de ce texte, en comparant chacun de ces mots conceptuels à ceux du dictionnaire dans lequel les mots sont chacun associés à une position dans le repère $\mathfrak{R}$.

En cas d'identité entre un mot conceptuel du texte et un mot du dictionnaire, on lit dans le dictionnaire la position, dans le repère $\Re$, associée à ce mot et on attribue cette position au mot conceptuel du texte. Les positions ainsi déterminées des mots conceptuels extraits des textes 1 à 3 sont celles indiquées dans le tableau 1.

Puis, pour chaque texte à archiver, on détermine la résultante des positions dans le repère $\Re$ de tous les mots conceptuels du texte, en multiplexant ces positions par un algorithme de composition. Celui-ci consiste ici à faire la somme vectorielle des positions de tous les mots conceptuels du texte à archiver, c'est-à-dire à additionner les coordonnées homologues des positions des mots conceptuels du texte.

Puis on normalise la résultante des positions de tous mots conceptuels du texte à archiver, et on obtient alors la position d'une conceptualisation globale de ce texte dans le repère $\Re$.

Par définition, un vecteur est normalisé lorsque sa norme est égale à un. L'étape visant à "normaliser" un vecteur consiste donc à diviser ce vecteur par sa propre norme.

La formule mathématique pour la détermination de la position de conceptualisation globale du texte d'indice j est donc:

$$\vec{t}_j = \frac{\vec{T}_j}{\left\|\vec{T}_j\right\|} = \frac{\sum_{i=1}^{N_j} \vec{m}_{ij}}{\left\|\sum_{i=1}^{N_j} \vec{m}_{ij}\right\|}$$

- $\vec{m}_{ij}$ représente le vecteur du mot conceptuel d'indice i du texte d'indice j,
- $\vec{T}_j$ représente la résultante des positions de tous les mots conceptuels du texte d'indice j et
- $\vec{t}_j$ représente le vecteur de conceptualisation globale du texte d'indice j,

avec i entier naturel variant de 1 à N_j (N_j représentant le nombre total de mots conceptuels du texte d'indice j), et j entier naturel variant de 1 à 3.

Le vecteur $\vec{t}_j$ de conceptualisation globale du texte d'indice j constitue une représentation vectorielle, dans le repère conceptuel $\Re$, du sens global du texte d'indice j.

Les coordonnées des vecteurs $\vec{t}_1$, $\vec{t}_2$, $\vec{t}_3$ de conceptualisation globale des textes 1, 2 et 3, respectivement, sont répertoriées dans le tableau 3.

Tableau 3

Texte j	Résultante $\vec{T}_j$	Vecteur de conceptualisation globale $\vec{t}_j$
Texte 1	(3, 2, 1)	(0.802, 0.535, 0.267)
Texte 2	(0, 1, 1)	(0, 0.707, 0.707)
Texte 3	(2, 1, 1)	(0.816, 0.408, 0.408)

Enfin, on archive les positions de conceptualisation globale des textes 1, 2 et 3.

2- Recherche, parmi la pluralité de textes archivés, de ceux qui traitent d'une question

On souhaite maintenant rechercher, parmi les textes archivés (textes 1, 2 et 3), les textes qui traitent d'une question déterminée qui est ici "la mécanique des fluides".

Comme pour l'archivage de texte, on procède à une analyse syntaxique des mots de la question afin d'en extraire les mots conceptuels qui sont ici "mécanique" et "fluide".

Dans le cas où la question comprendrait des mots conceptuels fléchis, on pourrait transformer ces mots dans leur forme non fléchie.

On compare chacun des mots conceptuels de la question à ceux du dictionnaire afin de déterminer leur position dans le repère conceptuel $\Re$. Les positions respectives du mot "mécanique" et du mot "fluide" sont indiquées dans le tableau 1.

Puis on détermine la résultante $\vec{Q}$ des positions de tous les mots conceptuels de la question, en multiplexant les positions des mots conceptuels de la question par l'algorithme de composition utilisé pour

l'archivage de textes. Enfin, on normalise la résultante $\vec{Q}$ afin d'obtenir le vecteur $\vec{q}$ de conceptualisation globale de la question.

Les vecteurs $\vec{Q}$ et $\vec{q}$ ont respectivement pour coordonnées (2, 1, 0) et (0.894, 0.447, 0).

Puis, on compare la position de la conceptualisation globale de la question aux positions homologues, de conceptualisation globale, des textes archivés pour retenir au moins l'une d'entre elles, correspondant à un texte recherché. Cette comparaison consiste à calculer, pour chaque texte archivé d'indice j (avec j entier naturel égal à 1, 2 ou 3), la distance D_j entre les deux positions respectives de la question et du texte.

La distance D_j entre le vecteur $\vec{q}$ de conceptualisation globale de la question et le vecteur $\vec{t}_j$ de conceptualisation globale du texte archivé d'indice j est ici calculée à l'aide de la formule suivante:

$$D_j = 1 - \langle \vec{t}_j, \vec{q} \rangle$$

On soulignera que le calcul de la distance D_j utilise le produit scalaire entre le vecteur $\vec{t}_j$ du texte d'indice j et le vecteur $\vec{q}$ de la question ($<\vec{t}_j,\vec{q}>$).

Le calcul de la distance D_j entre les positions respectives de la question et de chacun des textes archivés d'indice j (avec j égal à 1, 2 ou 3) permet d'évaluer la ressemblance entre la question et chacun des textes archivés.

Les résultats de ces calculs de distance sont indiqués dans le tableau 4.

Tableau 4

	Distance D_j
texte 1 / question	0, 044
texte 2 / question	0,688
texte 3 / question	0,088

D'après ces résultats, le texte le plus pertinent, qui est celui pour lequel la distance D_j est la plus faible, est le texte 1, ce qui correspond bien à la réalité.

On soulignera que le texte 1 est déterminé plus pertinent que le texte 3, malgré la présence dans ce dernier du terme "fluide".

Dans la description qui précède, le vecteur de conceptualisation globale, d'un texte ou de la question, est la résultante normalisée des positions de tous les mots conceptuels, de ce texte ou de la question. On pourrait également envisager de définir le vecteur de conceptualisation globale, d'un texte ou d'une question, comme la résultante non normalisée des positions de tous les mots conceptuels, de ce texte ou de cette question.

La formule pour le calcul de la distance D_j entre les positions respectives de la question et d'un texte archivé d'indice j serait alors la suivante:

$$D_j = 1 - \frac{\langle \vec{Q}, \vec{T}_j \rangle}{\|\vec{Q}\| \cdot \|\vec{T}_j\|}$$

- $\vec{Q}$ représente le vecteur de conceptualisation globale de la question et
- $\vec{T}_j$ représente le vecteur de conceptualisation globale du texte d'indice j.

En fait, dans ce cas, on normalise la résultante des positions des mots conceptuels par le calcul de la distance entre les positions respectives de conceptualisation globale du texte et de la question.

Dans une variante, ne différant de la description précédemment explicitée que par ce qui va maintenant être décrit, on munit l'espace vectoriel multidimensionnel d'un produit scalaire non euclidien et d'une norme associée non euclidienne.

On définit le produit scalaire non euclidien, de deux vecteurs $\vec{X}$ et $\vec{Y}$, par la formule suivante:

$$\langle \vec{X}, \vec{Y} \rangle = \sum_{i=1}^{n} \frac{1}{k_i} \cdot x_i \cdot y_i$$

On définit la norme associée du vecteur $\vec{X}$ par la formule suivante:

$$\|\vec{X}\| = \sqrt{\sum_{i=1}^{n} \frac{1}{k_i} \cdot x_i^2}$$

- x_i et y_i représentent les coordonnées respectives du vecteur $\vec{X}$ et du vecteur $\vec{Y}$ suivant l'axe A_i du repère conceptuel et
- k_i représente un coefficient de pondération, relatif à l'axe A_i,

avec i entier naturel variant de 1 à n, n représentant la dimension de l'espace vectoriel.

On fixe le coefficient k_i relatif à l'axe d'indice i en fonction de l'importance du concept exprimé par cet axe dans le repère conceptuel.

Dans cette variante, pour rechercher, parmi une pluralité de textes archivés, ceux qui sont pertinents à l'égard d'une question, on compare les positions des conceptualisations globales de la question et des textes archivés, en déterminant, pour chaque texte, la distance entre les deux positions respectives de la question et du texte, à l'aide de la formule de calcul de distance explicitée dans le premier mode de réalisation du procédé de recherche décrit, et en utilisant le produit scalaire non euclidien tel que défini ci-dessus.

Dans un deuxième mode de réalisation du procédé d'archivage de textes, ne différant du premier mode de réalisation décrit que par ce qui va maintenant être décrit, pour chaque texte à archiver, on associe à la position $P_{\Re}$, dans le repère $\Re$, de chaque mot conceptuel de ce texte d'abord sa position dans le texte P_T ainsi que son rôle syntaxique R_{synt} dans le texte, afin de former, pour chaque mot conceptuel extrait du texte, un triplet ($P_{\Re}$, P_T, R_{synt}) contenant la position $P_{\Re}$,dans le repère $\Re$, du mot, sa position P_T dans le texte et son rôle syntaxique R_{synt}.

Pour chaque texte à archiver, on détermine la résultante des positions des mots conceptuels du texte, en multiplexant les triplets de tous les mots consptuels du texte par un algorithme de composition, afin de déterminer la position de la conceptualisation globale de ce texte.

Pour rechercher, parmi les textes archivés suivant ce procédé d'archivage, ceux qui traitent d'une question, on détermine la position de la conceptualisation globale de la question. Pour cela, comme pour l'archivage des textes, on détermine la résultante des positions de mots conceptuels de la question en associant chaque mot conceptuel de la question à un triplet contenant la position de ce mot dans le repère $\Re$, sa position dans la question et son rôle syntaxique dans la question et en multiplexant ces triplets par l'algorithme de composition utilisé pour l'archivage.

Puis, on compare la position de la conceptualisation globale de la question aux positions homologues des textes archivés, en calculant la distance entre ces positions. On en déduit la ressemblance entre la question et les textes archivés et, par conséquent, les textes les plus pertinents qui traitent de la question.

Dans un troisième mode de réalisation du procédé d'archivage de textes, ne différant du premier mode de réalisation décrit que par ce qui va maintenant être décrit, on découpe le texte en une pluralité de segments. Chaque segment comprend initialement un nombre prédéfini de mots conceptuels, ici cinq, voisins l'un de l'autre dans le texte.

Deux segments sont dits "voisins" ici lorsqu'ils sont côte à côte dans le texte ou séparés l'un de l'autre uniquement par des mots non conceptuels.

On détermine les positions, dans le repère conceptuel, de tous les mots conceptuels du texte. Pour chaque segment de texte, on détermine la résultante des positions de tous les mots conceptuels de ce segment, en multiplexant ces positions par l'algorithme de composition utilisé dans le premier mode de réalisation du procédé d'archivage décrit. Puis on normalise cette résultante afin d'obtenir la position de conceptualisation globale du segment dans le repère conceptuel.

On compare ensuite deux à deux les positions de conceptualisation globale des segments voisins dans le texte, en calculant, pour chaque couple de segments voisins, la distance entre les deux positions respectives de conceptualisation des deux segments, à l'aide de la formule de calcul de la distance explicitée dans le premier mode de réalisation du procédé de recherche.

Si la distance entre les positions respectives de conceptualisation globale de deux segments voisins est inférieure à un seuil prédéfini, en d'autres termes si ces deux segments ont des sens proches, on regroupe ces deux segments en formant ainsi un nouveau segment dont on détermine la position de conceptualisation globale.

En revanche, si la distance entre les positions de conceptualisation globale de deux segments voisins est supérieure au seuil prédéfini, autrement dit si

ces deux segments ont des sens éloignés, on ne regroupe pas les deux segments.

On réitère l'étape consistant à regrouper les segments voisins, jusqu'à ne plus pouvoir les regrouper. Par regroupements itératifs de segments, on forme et on délimite ainsi une pluralité de portions de texte qui sont telles que la distance entre les positions respectives de conceptualisation globale de deux portions de texte voisines est supérieure au seuil prédéfini. En d'autres termes, le sens global de chaque partie du texte est éloigné du sens global d'une partie voisine.

Pour comparer une question et un texte archivé comprenant une pluralité de portions représentées chacune par sa position de conceptualisation globale dans le repère conceptuel, on compare la position de chacune des portions de texte à celle de la question, en calculant la distance entre ces positions. On considère un texte comme pertinent si la distance entre la position de l'une de ses portions et la position de la question est faible.

Bien entendu, on pourrait découper la question en une pluralité de portions représentées chacune par sa position de conceptualisation globale.

Dans ce cas, on comparerait deux à deux les vecteurs des portions d'un texte archivé et ceux des portions de la question. On considère que le texte est pertinent si la distance entre la position de l'une de ses portions et la position de l'une des portions de la question est faible.

On soulignera que, dans le troisième mode de réalisation du procédé d'archivage, on archive chacune des portions d'un texte de la même manière que l'on archive un texte (constitué d'une seule portion) dans le premier mode du procédé d'archivage. Finalement, un "texte" et une "portion de texte" sont deux ensembles de mots équivalents.

Concernant l'algorithme de composition pour la détermination de la résultante des positions de mots conceptuels d'un texte, d'un segment de texte ou d'une question, au lieu de faire seulement la somme vectorielle des positions des mots conceptuels du texte, du segment de texte ou de la question, on pourrait en outre amplifier les valeurs des coordonnées les plus fortes du vecteur résultant de la somme vectorielle des positions des mots conceptuels, par exemple en les multipliant par un coefficient prédéfini. On amplifie ainsi encore l'importance des concepts les plus

importants, au détriment des concepts moins importants, afin d'éviter d'éventuelles ambiguïtés lors de la comparaison des vecteurs de conceptualisation globale d'un texte et d'une question. En fait, on réduit ainsi le bruit dû aux coordonnées ayant des valeurs faibles des vecteurs de conceptualisation.

Afin d'illustrer cette variante, appliquons la au texte 1. Par la somme vectorielle des positions de tous les mots conceptuels de ce texte, on obtient le vecteur (3, 2, 1). Afin d'obtenir la résultante des positions de tous les mots conceptuels du texte 1, on multiplie les coordonnées les plus fortes, qui sont celles suivant les axes A_1 et A_2, par un coefficient ici égal à 2. La résultante du texte 1 est donc le vecteur (6, 4, 1).

Dans l'exemple didactique décrit plus haut, la question, "la mécanique des fluides", comprenait peu de mots. On pourrait bien évidemment prendre une question contenant beaucoup plus de mots et consistait même en un texte.

En pratique, le repère conceptuel $\mathfrak{R}$ comprend plusieurs centaines de dimensions, et le dictionnaire contient plusieurs milliers de mots.

REVENDICATIONS

1- Procédé d'archivage d'un texte (1) selon lequel:
- on crée dans un repère conceptuel multidimensionnel un dictionnaire de mots,
- on compare chaque mot conceptuel d'au moins une portion du texte à archiver (1) à ceux du dictionnaire pour déterminer la position de ce mot dans ledit repère et
- on détermine la résultante ($\vec{T}_1$) des positions de tous les mots conceptuels de la portion de texte à archiver (1) pour déterminer la position d'une conceptualisation globale de la portion de texte (1) dans ledit repère et archiver cette position.

2- Procédé selon la revendication 1, dans lequel, pour déterminer la résultante des positions, dans le repère, de tous les mots conceptuels de la portion de texte à archiver, à chaque position de mot dans le repère, on associe d'abord sa position dans le texte et son rôle syntaxique.

3- Procédé selon la revendication 1, dans lequel pour déterminer la résultante ($\vec{T}_1$) des positions des mots conceptuels de la portion de texte à archiver (1), on multiplexe ces positions par un algorithme de composition.

4- Procédé selon la revendication 3, dans lequel l'algorithme de composition consiste à faire la somme vectorielle des positions de tous les mots conceptuels de la portion de texte à archiver (1).

5- Procédé selon la revendication 4, dans lequel l'algorithme de composition consiste en outre à amplifier l'importance des concepts les plus importants.

6- Procédé selon la revendication 1, dans lequel on normalise la résultante ($\vec{T}_1$) des positions de tous les mots conceptuels de la portion de texte à archiver (1).

7- Procédé selon la revendication 1, dans lequel on rend le repère conceptuel multidimensionnel orthonormé.

8- Procédé selon la revendication 1, dans lequel, pour chaque mot à introduire dans le dictionnaire, on recherche tous les concepts relatifs au repère conceptuel, auxquels ce mot est susceptible de faire référence et, suivant ces concepts, on attribue au mot une position dans le repère conceptuel.

9- Procédé selon la revendication 1, dans lequel on procède à une analyse syntaxique de tous les mots de la portion de texte (1) afin d'en extraire les mots conceptuels.

10- Procédé selon la revendication 1, dans lequel on transforme les mots fléchis de la portion de texte à archiver (1) dans leur forme non fléchie.

11- Procédé d'archivage d'un texte comprenant une pluralité de portions de texte, dans lequel on archive chaque portion de texte selon le procédé de la revendication 1.

12- Procédé selon la revendication 11, dans lequel on découpe le texte en une pluralité de segments dont on détermine les positions respectives de conceptualisation globale dans le repère conceptuel, et on compare les positions respectives de conceptualisation globale des segments voisins dans le texte pour délimiter les portions du texte.

13- Procédé selon la revendication 11, dans lequel pour comparer les positions respectives de conceptualisation globale de deux segments voisins dans le texte, on détermine la distance entre ces positions, et, dans le cas où ladite distance est inférieure à un seuil prédéfini, on regroupe les deux segments en formant un nouveau segment.

14- Procédé selon la revendication 13, dans lequel on forme les portions de texte par regroupements itératifs de segments.

15- Procédé de recherche parmi une pluralité de textes archivés selon le procédé d'archivage de la revendication 1, de ceux qui traitent d'une question, dans lequel:
- comme pour l'archivage de texte, on détermine la position dans le repère conceptuel multidimensionnel d'une conceptualisation globale de la question, par détermination de la résultante ($\vec{Q}$) des positions de tous les mots conceptuels de la question et

- on compare la position de la conceptualisation globale de la question aux positions homologues des textes archivés, pour retenir au moins l'une d'entre elles, correspondant à un texte recherché.

16- Procédé selon la revendication 15, dans lequel on compare les positions des conceptualisations globales de la question et des textes archivés en déterminant, pour chaque texte, la distance entre les deux positions respectives de la question et du texte.

17- Procédé selon la revendication 15, dans lequel le calcul de la distance entre deux positions dans le repère conceptuel utilise le produit scalaire desdites positions.

18- Procédé selon la revendication 17, dans lequel on calcule la distance entre deux positions dans le repère conceptuel, à l'aide de la formule suivante:

$$D = 1 - \frac{\langle \vec{X}, \vec{Y} \rangle}{\|\vec{X}\| \cdot \|\vec{Y}\|}$$

- $\vec{X}$ et $\vec{Y}$ représentant les deux positions,
- D représentant la distance entre les deux positions $\vec{X}$ et $\vec{Y}$,
- $<\vec{X}, \vec{Y}>$ représentant le produit scalaire de $\vec{X}$ et de $\vec{Y}$ et
- $\|\vec{X}\|$ et $\|\vec{Y}\|$ représentant les normes respectives de $\vec{X}$ et de $\vec{Y}$.

19- Procédé selon la revendication 15, dans lequel la distance déterminée entre deux positions est non euclidienne.

20- Procédé selon la revendication 19, dans lequel la distance déterminée entre deux positions utilise le produit scalaire défini par la formule suivante:

$$\langle \vec{X}, \vec{Y} \rangle = \sum_{i=1}^{n} \frac{1}{k_i} \cdot x_i \cdot y_i$$

- $<\vec{X}, \vec{Y}>$ représentant le produit scalaire de deux positions $\vec{X}$ et $\vec{Y}$,
- n, entier naturel, représentant la dimension du repère conceptuel comportant n axes d'indice i avec i entier naturel variant de 1 à n,
- x_i et y_i représentant les coordonnées respectives des positions X et Y suivant l'axe d'indice i et
- k_i représentant un coefficient de pondération relatif à l'axe d'indice i.

21- Procédé selon la revendication 15, dans lequel on normalise la résultante ($\vec{Q}$) des positions de tous les mots conceptuels de la question.

22- Procédé selon la revendication 15, dans lequel on procède à une analyse syntaxique de tous les mots de la question afin d'en extraire les mots conceptuels.

24- Procédé selon la revendication 15, dans lequel on transforme les mots fléchis de la question dans leur forme non fléchie.

ABREGE

PROCEDE D'ARCHIVAGE DE TEXTES ET PROCEDE DE RECHERCHE, PARMI LES TEXTES ARCHIVES, DE CEUX QUI SONT PERTINENTS A L'EGARD D'UNE QUESTION

Procédé d'archivage: on crée dans un repère conceptuel multidimensionnel un dictionnaire de mots, on compare chaque mot conceptuel d'au moins une portion du texte à archiver à ceux du dictionnaire pour déterminer la position de ce mot dans ledit repère et on détermine la résultante (T_1) des positions de tous les mots conceptuels de la portion de texte à archiver pour déterminer la position d'une conceptualisation globale de la portion de texte dans ledit repère et archiver cette position.

Procédé de recherche: on détermine la position dans le repère conceptuel multidimensionnel d'une conceptualisation globale de la question, et on compare la position de la conceptualisation globale de la question aux positions homologues des textes archivés, pour retenir au moins l'une d'entre elles, correspondant à un texte recherché.

Figure unique